Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **513,** 2509–2521 (2022)
Advance Access publication 2022 April 20

https://doi.org/10.1093/mnras/stac1007

Tracing Milky Way substructure with an RR Lyrae hierarchical clustering forest

Brian T. Cook [1,2★†] Deborah F. Woods,[2] Jessica D. Ruprecht,[2] Jacob Varey,[2] Radha Mastandrea,[2,3] Kaylee de Soto,[3,2] Jacob F. Harburg,[2] Umaa Rebbapragada[4] and Ashish A. Mahabal [5,6]

[1] *Center for Relativistic Astrophysics, School of Physics, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[2] *MIT Lincoln Laboratory, Lexington, MA 02421, USA*
[3] *Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02142, USA*
[4] *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA*
[5] *Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*
[6] *Center for Data Driven Discovery, California Institute of Technology Pasadena, CA 91125, USA*

Accepted 2022 April 7. Received 2022 April 6; in original form 2021 July 22

ABSTRACT

RR Lyrae variable stars have long been reliable standard candles used to discern structure in the Local Group. With this in mind, we present a routine to identify groupings containing a statistically significant number of RR Lyrae variables in the Milky Way environment. RR Lyrae variable groupings, or substructures, with potential Galactic archaeology applications are found using a forest of agglomerative, hierarchical clustering trees, whose leaves are Milky Way RR Lyrae variables. Each grouping is validated by ensuring that the internal RR Lyrae variable proper motions are sufficiently correlated. Photometric information was collected from the *Gaia* second data release and proper motions from the (early) third data release. After applying this routine to the catalogue of 91 234 variables, we are able to report 16 unique RR Lyrae substructures with physical sizes of less than 1 kpc. Five of these substructures are in close proximity to Milky Way globular clusters with previously known tidal tails and/or a potential connection to Galactic merger events. One candidate substructure is in the neighbourhood of the Large Magellanic Cloud but is more distant (and older) than known satellites of the dwarf galaxy. Our study ends with a discussion of ways in which future surveys could be applied to the discovery of Milky Way stellar streams.

Key words: catalogues – stars: variables: RR Lyrae – Galaxy: halo – Galaxy: stellar content – Galaxy: structure.

1 INTRODUCTION

The Milky Way's stellar population is laced with substructures from which dwarf galaxy merger events and star formation histories can be inferred across cosmic time. The hierarchical galaxy formation picture suggests that galaxies like our own grow with time as orbiting dwarves are accreted (e.g. Press & Schechter 1974; Bullock & Johnston 2005). Tidal shocking and subsequent stripping occur during these mergers (see Binney & Tremaine 2008, for a review); however, disrupted clusters have been successfully identified using stellar configurations in the Milky Way's phase space (e.g. Portegies Zwart 2009; Malhan et al. 2019).

Many Galactic archaeological artefacts can be analysed using RR Lyrae variable stars, a standard candle popular in Local Group studies (see Preston 1964, for a review). Globular clusters (GCs), structures of Population II stars that formed very early in a galaxy's lifetime (Searle & Zinn 1978), are especially critical. The age and metallicity distributions of GCs, for example, can be used to trace the star formation of their host galaxy (Kissler-Patig 1999). GCs can also be helpful in determining the Milky Way's merger history (e.g. Kruijssen et al. 2020; Bonaca et al. 2021). RR Lyrae variables are most frequently associated with GCs (e.g. Zinn 1985; Clement et al. 2001), but their utility has been demonstrated in a variety of Milky Way satellite studies as well. The so-called Pisces Overdensity, a metal-poor satellite related to the infall of the Large Magellanic Cloud (LMC), has been successfully analysed using an RR Lyrae variable catalogue (e.g. Watkins et al. 2009; Belokurov et al. 2019).

RR Lyrae variables have also been found in stellar streams (e.g. Duffau et al. 2006; Vivas et al. 2016), showing that this variable star class populates Galactic substructure in the form of tidal debris. This property is useful even after the progenitor object has undergone serious tidal distortion. The kinematic space morphology of debris from infalling dwarves can, in turn, be used to make inferences about the Galactic merger history. There are three general tidal debris morphology classes, with early merger events manifesting primarily as 'cloudy' and 'great circle' debris (Johnston et al. 2008), the former comprised mostly of stars on highly eccentric orbits. A recent study of RR Lyrae variables classified using *Gaia* DR2 (Iorio & Belokurov 2019) does not use this classification system, but confirms that RR

★ E-mail: briantcook3070@gmail.com
† Distribution A. Approved for public release - distribution is unlimited.

Lyrae variables in the inner Galactic halo can be classified as 'cloudy' debris that were primarily sourced by a single ancient merger event.

Thus, we are motivated to find substructures in the Milky Way environment comprised of RR Lyrae variable stars, and there is a suite of clustering methodologies to be considered. The easiest set of objects to cluster would have obvious categorical features; if 20 people were distributed between New York, Tokyo, and Buenos Aires, for example, then there would be three clusters found within a data structure tabulating personal separation distances, and we would probably not need to write a computer program to determine the clusters. The distribution of Milky Way RR Lyrae variables, however, is remarkably more complex, so we must rely upon an automated clustering algorithm to glean RR Lyrae substructures. One of our primary goals is to determine an optimal clustering method such that rich structure can be found without overfitting to the data. One potential method is a Gaussian Mixture Model that incorporates an overfitting penalty (such as the AIC, see Akaike 1998), but this would be built upon the assumption that the variable stars are organized into subpopulations that are independently normally distributed. Choosing a number of clusters or nearest neighbours a priori and then iteratively optimizing a cost function is another option, but this would provide no information as to what extent neighbouring clusters are related.

An agglomerative, hierarchical clustering algorithm, when applied to a catalogue of RR Lyrae variables, mitigates both of these concerns. Agglomerative, hierarchical clustering is the process by which n clusters each containing one object (in our case, an RR Lyrae variable 'leaf') accumulate in a branch-like fashion into a single 'root' cluster populated by n objects (Gower & Ross 1969; Gordon 1987; Everitt et al. 2011). Hierarchical clustering has been applied in other astronomical contexts, including galaxy distributions (Peebles 1974) and analyses of the SDSS 'Great Wall' (Ivezić et al. 2014). Analyses of hierarchical structures of this kind are demonstrably a powerful method of interpreting a variety of complex networks (Clauset, Moore & Newman 2008). It is worth noting, however, that using the term 'clusters' here is an unfortunate coincidence, as we do not want to unintentionally conflate our findings with open and GCs. Confirming a single merger via analyses of the tree's root (i.e. RR Lyrae distribution on the largest scales) is also beyond the scope of this study, as we will discuss in Section 2. It is encouraging, none the less, that hierarchical clustering trees might help connect the Milky Way's RR Lyrae population to Galactic merger events in a variety of ways.

This study's contributions to the Galactic archaeology literature is twofold: we present a new way of identifying substructures in the Milky Way using RR Lyrae hierarchical clustering trees (described in Section 2.3) whose coverage includes the entire sky plane, as well as substructure candidates suitable for future studies, including a previously unknown mid-halo substructure and potential LMC satellite. The standardization of photometric and proper motion information, as well as our clustering algorithm, is described and tested in Section 2. Each RR Lyrae variable is placed in a five-dimensional phase space (distance modulus + 2D sky plane coordinates + 2D sky plane proper motions) used for clustering and validation. We explore a small portion of the stellar stream parameter space to test our algorithm's effectiveness, and confirm its utility by identifying a known stream of RR Lyrae variables. Our method of identifying RR Lyrae variable substructures and the subsequent results are presented in Section 3. A discussion in Section 4 focuses on the relationship between forest groupings and the Milky Way GC population, as well as potential future studies in which kinematic and spectroscopic information could be leveraged to identify Galactic stellar streams.

2 METHODS

2.1 RR Lyrae variables as standard candles

A well-worn rung on the cosmic distance ladder, as mentioned in Section 1, is the RR Lyrae variable star. The variability of the RR Lyrae is due to changes in the stellar opacity, which causes a periodic ebb and flow of the star's luminosity (Maeder 2009). The period of the RR Lyrae variable directly relates to its absolute magnitude, allowing astronomers to calculate absolute distances to stars of this type. Once the absolute magnitude has been computed from the variable's pulsation period, the luminosity can be inferred using the solar luminosity and bolometric flux:

$$L_\star = L_\odot \, 10^{0.4\left[M_{\mathrm{bol},\odot} - (M_\mathrm{V} + \mathrm{BC})\right]}, \tag{1}$$

where BC is the bolometric correction applied to the absolute magnitude of the RR Lyrae variable in the Johnson V band; see Sandage & Cacciari (1990) for a table of RR Lyrae bolometric corrections. The luminosity distance can then be computed with an observed (mean) brightness $b_\star$. However, absolute magnitudes can only be inferred from the period in infrared bands (Catelan, Pritzl & Smith 2004); consequently, we must rely upon other information to determine a particular variable's absolute magnitude when considering visible wavelength data. Chaboyer (1999) provides a relation between the absolute magnitude of RRab variables and the metallicity ([Fe/H]):

$$M_\mathrm{V} = (0.23 \pm 0.04)\,[\mathrm{Fe/H}] + (0.93 \pm 0.12). \tag{2}$$

RR Lyrae variables belong to one of three classes based on the shape of their light curve (Smith 1995); RRab variables are the most common, and we restrict our catalogue of RR Lyrae variables to those of type RRab. The terms RR and RRab Lyrae variable stars will be used interchangeably throughout this paper.

The distance modulus $\mu_\mathrm{RRL} \equiv m_\mathrm{V} - M_\mathrm{V} - A_\mathrm{V}$, where m_V, M_V are the apparent and absolute magnitudes in the Johnson V band, and A_V is the associated extinction, i.e. reddening due to dust and other intermediate material, provides a well-established logarithmic scale for stellar distances:

$$\mu_\mathrm{RRL} = 10 + 5\left(\log_{10} \tfrac{d_\mathrm{RRL}}{[1\,\mathrm{kpc}]}\right). \tag{3}$$

The normalized RR Lyrae distance modulus uncertainty can be determined using error propagation:

$$\delta\mu_\mathrm{RRL} = \sqrt{\delta m_\mathrm{V}^2 + \delta M_\mathrm{V}^2 + A_\mathrm{V}^2}, \tag{4}$$

$$\epsilon_{\mu_\mathrm{RRL}} \equiv \tfrac{1}{\mu_\mathrm{RRL}}\,\delta\mu_\mathrm{RRL}, \tag{5}$$

where the uncertainty of a measured value x is denoted δx. Equation (4) establishes the relationship between the measured distance modulus uncertainty and absolute magnitude uncertainty. We proceed with using the distance modulus as a proxy for physical distance, as the associated uncertainty is typically an order-of-magnitude smaller. In order to ensure that we avoid identifying specious structures at large distance moduli, we set an upper limit on substructure size in physical space and then translate to an appropriate distance moduli spread $\Delta\mu_\mathrm{RRL}$.

2.2 RRab variable data from the *Gaia* mission

RR Lyrae variable stars have been identified using the *Gaia* mission's Specific Object Study (SOS) pipeline (Clementini et al. 2019). The SOS pipeline ingested DR2 time-series photometry from the second data release (DR2) in the *Gaia* multibands (G, G_{BP}, and G_{RP}).

Processing was predicated upon the period–amplitude and period–luminosity relations found within each variable star's G-band light curve, as well as colour–magnitude relations reported in DR2. Each variable star's best classification was then reported as part of DR2; for this study, we retained variables of type RRab for further analysis.

The translation between the Johnson V band and *Gaia* multibands can be approximated in the following way:

$$m_{\rm V} = G - \textstyle\sum_{n=0}^{2} a_n (G_{BP} - G_{RP})^n, \tag{6}$$

$$\delta m_{\rm V} = \sqrt{\delta G^2 + (a_1 + 2a_2(G_{BP} - G_{RP}))^2 (\delta G_{BP}^2 + \delta G_{RP}^2)}, \tag{7}$$

where $\{a_n\}$ are polynomial coefficients fit to *Gaia* DR2 data.[1] In cases where the $(G_{BP} - G_{RP})$ colours or blue/red-band uncertainties were unavailable in the *Gaia* DR2 catalogue, we used, if available in the *Gaia* EDR3 catalogue, error propagation from the mean fluxes in each band with their associated uncertainties to compute $G_{BP} - G_{RP}$, δG_{BP}, and δG_{RP}. The reported (or inferred) *Gaia* magnitudes, and their uncertainties, could be used to compute $(m_{\rm V}, \delta m_{\rm V})$. If there were insufficient data available to make such inferences, the RRab variable datum was discarded.

The absolute magnitude and its associated uncertainty can be determined using equation (2):

$$\delta M_{\rm V} = 0.23\,(\delta[{\rm Fe/H}]) + 0.04\,\big|[{\rm Fe/H}]\big| + 0.12, \tag{8}$$

where δ[Fe/H] is the reported metallicity uncertainty. RRab variables without a known metallicity were assigned the median value of the well-defined RRab variable sample, with an associated uncertainty equal to half the total metallicity domain, $\delta[{\rm Fe/H}] = \frac{1}{2}[\max([{\rm Fe/H}]) - \min([{\rm Fe/H}])]$.

The extinction due to intermediate dust was computed using a publicly available data cube containing a 3D map of Milky Way dust; see Green (2018), Green et al. (2019) for more details, as well as Schlafly & Finkbeiner (2011) to see how reddening translates to extinction. This mapping provides an extinction vector along each sightline given as input, where each vector element symbolizes the extinction value $A_{\rm V}(\mu)$ at that particular distance modulus along the sightline. Each $(A_{\rm V}, \delta A_{\rm V})$ pair was computed using the following scheme for each RR Lyrae variable star; if the distance modulus was within the range of support, the median and standard deviation of the five nearest vector elements were collected. In the event the unreddened distance modulus of the star was beyond the reported range of support of a well-defined extinction vector, the extinction value took on the maximum (far-field) vector value and the uncertainty was left equal to zero. This choice was motivated by figs 1–5 in Green et al. (2019), where it is shown that the total extinction integrated to infinity is of the same order as the integrated extinction in the domain of the extinction vector $\mu \in (0, \mu_{\rm max} \sim 15]$, i.e.

$$\int_0^{\infty} \frac{{\rm d}A_{\rm V}}{{\rm d}\mu}\,{\rm d}\mu \simeq \int_0^{\mu_{\rm max}} \frac{{\rm d}A_{\rm V}}{{\rm d}\mu}\,{\rm d}\mu. \tag{9}$$

RR Lyrae variable stars whose corresponding extinction vector contained NaN values were assigned the following ordered pair values: $(\tilde{A}_{\rm V}, \frac{1}{2}[\max(A_{\rm V}) - \min(A_{\rm V})])$, where $\tilde{x}$ represents the median of a collection of x values.

In order to render out false positive stellar stream identifications from the hierarchical clustering trees we construct later on (see Section 2.3), we turn to the *Gaia* EDR3 catalogue (Gaia Collaboration 2021; Lindegren et al. 2021) for RR Lyrae variable star proper motions on the sky plane. The newest data release from the *Gaia* mission contains full astrometry for ∼1.5 billion sources, including parallaxes and proper motions. The *Gaia* cross-matching step is achieved using an ADQL query,[2] in which matches are returned if the source ID is in agreement between the DR2 and EDR3 catalogues. There are few radial velocities available from the *Gaia* DR2 catalogue available for these stars, so we restrict ourselves to only considering motion on the sky plane. We retained 91 234 RRab variables from the *Gaia* DR2 variable catalogue after the entire data collection and refinement routine, discarding data with missing or ill-defined values critical to the computation of the distance modulus. Fig. 1 shows that the majority of the RRab variables have an inferred distance modulus uncertainty $\lesssim 5$ per cent of the distance modulus itself, thus providing support for the usage of RR Lyrae variables in finding Milky Way substructures.

2.3 Particle-based, agglomerative, hierarchical clustering

A hierarchical clustering algorithm needs an $\mathbb{R}^{n\times n}$ matrix to encode the separation distances between the n objects that will populate the tree. Choosing the metric with which to calculate the distances depends on the nature of the objects; the standard metric is Euclidean, although distances in phase space and non-physical spaces that incorporate stellar attributes such as metallicity have been used to identify potential clusters in a variety of astronomical contexts (e.g. Maciejewski et al. 2009; De Silva et al. 2015). Using the catalogue data, we can construct three-dimensional vectors $\mathbf{x} \equiv (x, y, z)$ to get the distance between two variable stars with the Euclidean metric:

$$D(\mathbf{x}, \mathbf{x}') = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}, \tag{10}$$

$$x \equiv \mu_{\rm RRL} \cos\alpha \cos\delta, \tag{11}$$

$$y \equiv \mu_{\rm RRL} \sin\alpha \cos\delta, \tag{12}$$

$$z \equiv \mu_{\rm RRL} \sin\delta, \tag{13}$$

where $\mu_{\rm RRL}$ is the distance modulus. Each RR Lyrae variable's angular coordinates are expressed by its right ascension $(\phi = \alpha)$ and declination $(\theta = \delta)$. The distance between RR Lyrae variables in this metric have units of magnitude, which can be directly translated into a physical distance.

There are a number of choices in how to quantify the separation between two clusters, and for this analysis we have focused on the average linkage. The average linkage distance between clusters p and q is

$$L(p, q) = \frac{1}{n_p n_q} \sum_{i=1}^{n_p} \sum_{j=1}^{n_q} D(\mathbf{x}_{p,i}, \mathbf{x}_{q,j}), \tag{14}$$

where $D(\mathbf{x}_{p,i}, \mathbf{x}_{q,j})$ is the distance (as defined in equation 10) between the ith and jth variables in clusters p and q, respectively. This linkage method is a compromise between single linkage (minimum distance between points in compared clusters, susceptible to imbalances and 'chaining') and complete linkage (maximum distance between points in compared clusters, tends to generate clusters that are roughly uniform in size and shape).

The algorithm proceeds as follows: if we have n clusters $c_1, \ldots, c_n$, the $n(n-1)/2$ unique linkage distances are computed. Let us

[1] https://gea.esac.esa.int/archive/documentation/GDR2/Data_processing/chap_cu5pho/sec_cu5pho_calibr/ssec_cu5pho_PhotTransf.html

[2] https://gea.esac.esa.int/archive/

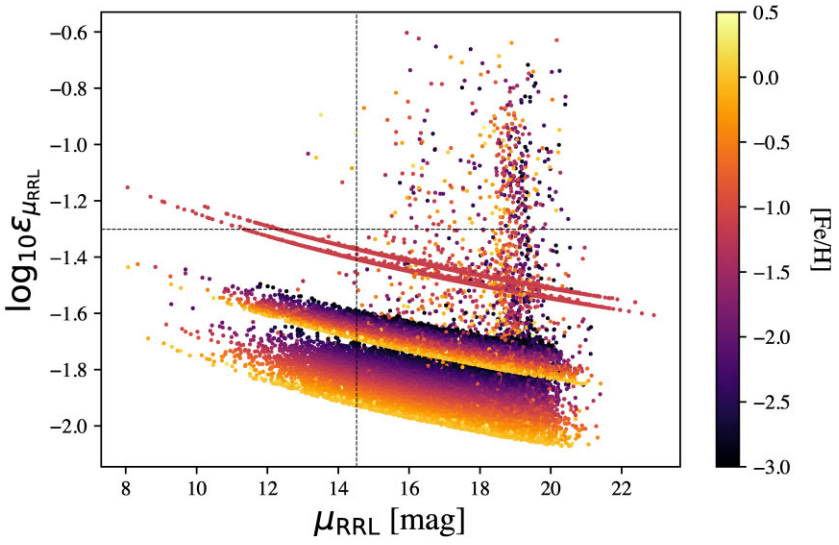


Figure 1. The distance modulus uncertainty $\epsilon_{\mu_{\rm RRL}}$ and distance modulus $\mu_{\rm RRL}$ of each RRab variable, where each data point is colour-coded by its reported or estimated metallicity. The dashed lines show the galactocentric distance (8 kpc) translated into distance modulus units, as well as the distance modulus uncertainty threshold consistent with 5 per cent of the computed distance modulus. A large number of RR Lyrae variables in the data set have a reported *Gaia* *G*-band value $G \simeq 19$, which explains the collection of variables with a distance modulus $\mu_{\rm RRL} \simeq 19$ that appears independent of the major $\log_{10}\epsilon_{\mu_{\rm RRL}}$ trends.

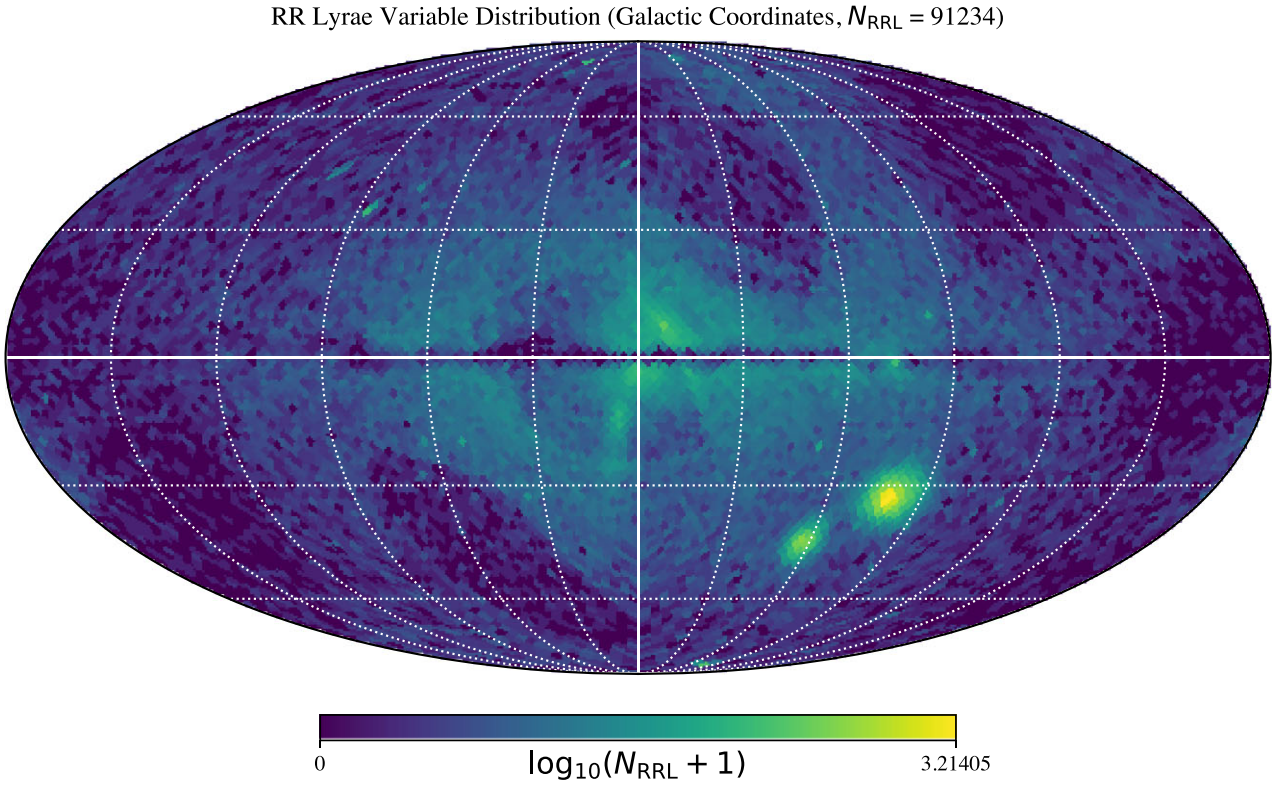


Figure 2. A HEALPix number density map ($N_{\rm pixels} = 12\,288$, see Section 3) of the RR Lyrae variables (type RRab) from the *Gaia* DR2 catalogue retained for this study. Notable regions with an overdensity of RR Lyrae variables include the Galactic centre and Magellanic Clouds. The projection is in Galactic coordinates (ℓ, b), where the grid is in 30 deg intervals.

say that the minimum linkage distance is between clusters c_i, c_j. The sets of RR Lyrae variables describing these two clusters are then combined into a single cluster c_{ij}, and at the next step there are $n-1$ clusters: $c_1, \ldots, c_{ij}, \ldots, c_{n-1}$. Once there is only a single cluster remaining, the algorithm terminates.

Agglomerative clustering does not provide a straightforward method for determining the optimal set of clusters, as the initial and final number of clusters are fixed. Intermediate steps can be analysed by 'cutting' the tree constructed from the condensed distance matrix at a specific height such that all identified clusters have an average

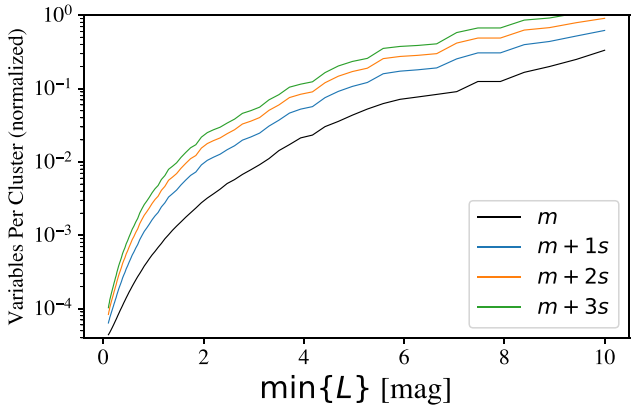


Figure 3. The mean number of RR Lyrae variables per cluster as a function of clustering scale, where the requisite hierarchical clustering tree was constructed from the catalogue subsample displayed in Fig. 4. Each cluster contains, on average, m RR Lyrae variables with a standard deviation s at each of the listed clustering scales. Clusters whose variable population exceeded $m + 3s$ were regarded as statistically significant.

linkage distance greater than or equal to that height. Langfelder, Zhang & Horvath (2007) provide a general method of tree-cutting that determines the best choice of clusters. If we cut the tree at a large number of heights and count the corresponding number of identified clusters at each height, we can identify a particular configuration that is resistant to further clustering. For example, if the minimum average linkage distance between n clusters is h_0 and the same minimum distance between $m - 1$ clusters is h_1, it stands to reason that the variables are neatly classified into n clusters if $|h_1 - h_0| \gg \Delta h$, where Δh is a tree-cutting height interval. Quantifying this height separation can be done by setting up an array of heights at which the tree is cut (separated by Δh), populating a corresponding array with the number of branches (clusters) at each height, and then computing the mode of that array to determine the optimal number of clusters. However, this procedure would then obscure interesting phenomena at the scales of GCs and stellar streams.

With agglomerative clustering, there is a suitable alternative; we can determine scales at which there are comparatively many clusters populated by a statistically significant number of RR Lyrae variables. Using the tree-cutting process, we compute the mean and standard deviation of cluster populations at each height. The heights at which the tree are cut are evenly spaced in log space between $h_{\rm min} = 0.1$ mag and $h_{\rm max} = 10$ mag; the motivation for this scheme is that we are interested in a deeper analysis of the stellar halo's substructure than in structures at Galactic scales.

Fig. 3 shows that at intermediate distance scales, the criterion for being categorized as a significantly populated cluster becomes more stringent. When the variables first begin to agglomerate ($\min\{L\} \sim 0.1$ mag, where $\{L\}$ is the set of linkage distances), there may be two or three variable stars that are considerably close; the mean number of variables per cluster is very close to 1 and the standard deviation is small in this regime, so this would be regarded as a cluster of interest. Such clusters should be approached with scepticism, given that the computed distance uncertainties are larger than these separation distances. On the largest scales ($\min\{L\} \gtrsim 10$), the standard deviation is considerably higher than the mean, so a cluster with a variable population $\approx N_{\rm RRL,\,total}$ would be the only significant cluster. If there are only two or three clusters remaining, however, then this would yield a high significant-to-insignificant-cluster ratio.

We collected all of the statistically significant clusters (i.e. $N_\star \geq m + 3s$, where m is the mean number of RR Lyrae variables per cluster and s is the associated standard deviation) at clustering scales ($N_{\rm cuts} = 80$ evenly separated in log space between $h_{\rm min}$ and $h_{\rm max}$ defined above) where the ratio of statistically significant to total clusters was maximized. Fig. 4 shows the results of clustering stars within a subregion of the sky plane, where stars belonging to statistically significant groupings are colour-coded by a grouping ID at a small clustering scale; we should expect to find groupings of RR Lyrae variables consistent with stellar streams at a variety of clustering scales, and the logarithmic spacing of our cuts ensures sampling of the trees' variety of branch sizes with a preference towards clustering scales where $\min\{L\} \lesssim 2$ mag. Collecting every cluster identified across all significant clustering scales would admit redundant groupings; from these redundancies, one was retained from each for further analysis while potential redundancies are listed in the Appendix (Section B).

The relevance of the identified groupings can be tested using *Gaia* EDR3 proper motions. Stellar streams moving on non-chaotic orbits through the Galactic tidal field will have well-correlated configurations in velocity space. This instructs our usage of velocity information in the form of sky plane proper motions. To start, we determine each grouping's median proper motion direction $\hat{v}$, written as a two-dimensional vector on the sky plane:

$$\hat{v} \equiv \frac{\langle \tilde{\mu}^*_\alpha, \tilde{\mu}_\delta \rangle}{\sqrt{\tilde{\mu}^{*^2}_\alpha + \tilde{\mu}^2_\delta}}, \tag{15}$$

where $\mu^*_\alpha \equiv \mu_\alpha \cos\delta$. These proper motion components are not to be confused with the distance modulus $\mu_{\rm RRL}$. We can determine the alignment of the ith star of the grouping by taking the dot product between its proper motion and the grouping's median proper motion:

$$\cos\theta_i = \hat{v} \cdot \hat{v}_i = \frac{(\tilde{\mu}^*_\alpha \mu^*_{\alpha,i} + \tilde{\mu}_\delta \mu_{\delta,i})}{\sqrt{\left(\tilde{\mu}^{*^2}_\alpha + \tilde{\mu}^2_\delta\right)\left(\mu^{*^2}_{\alpha,i} + \mu^2_{\delta,i}\right)}}. \tag{16}$$

The set

$$X_\theta \equiv \{\cos\theta_1, \ldots, \cos\theta_N\}, \tag{17}$$

which is a measure of how well each of the N stellar proper motions within the grouping are aligned with the median proper motion, will thus contain values ranging from -1 (antiparallel) to 1 (parallel). If the median of this set is below $\tilde{X}_{\theta,\rm min} = 0.98$, the grouping is discarded.

The proper motion validation layer is dependent on the following assumptions: the median proper motion vector is appropriate for comparison with all grouping members, and that the data within a grouping is only mildly heteroscedastic, i.e. that the statistical deviations from the median proper motion do not vary from one part of the grouping to the next. The first assumption breaks down for groupings distributed across large regions of the sky plane, and the second assumption breaks down for systems that have been disturbed by external objects like the Galactic Bar. The Pal 5 stellar stream has both of these qualities (Pearson, Price-Whelan & Johnston 2017), which explains why our routine can only identify the stream's core (see Section 2.4).

We only retain groupings with parameters similar to Sesar et al. (2013): $15 \leq N_{\rm RRL} \leq 40$ and $R_{\rm grouping} \leq R_{\rm max} = 1$ kpc; see Appendix A for motivation of the chosen $\tilde{X}_{\theta,\rm min}$, $R_{\rm max}$ values. The grouping size R translates to a (positive) distance modulus spread $\Delta\mu$ that is dependent on how far away the grouping is from the observer (i.e. distance modulus μ, equation 3):

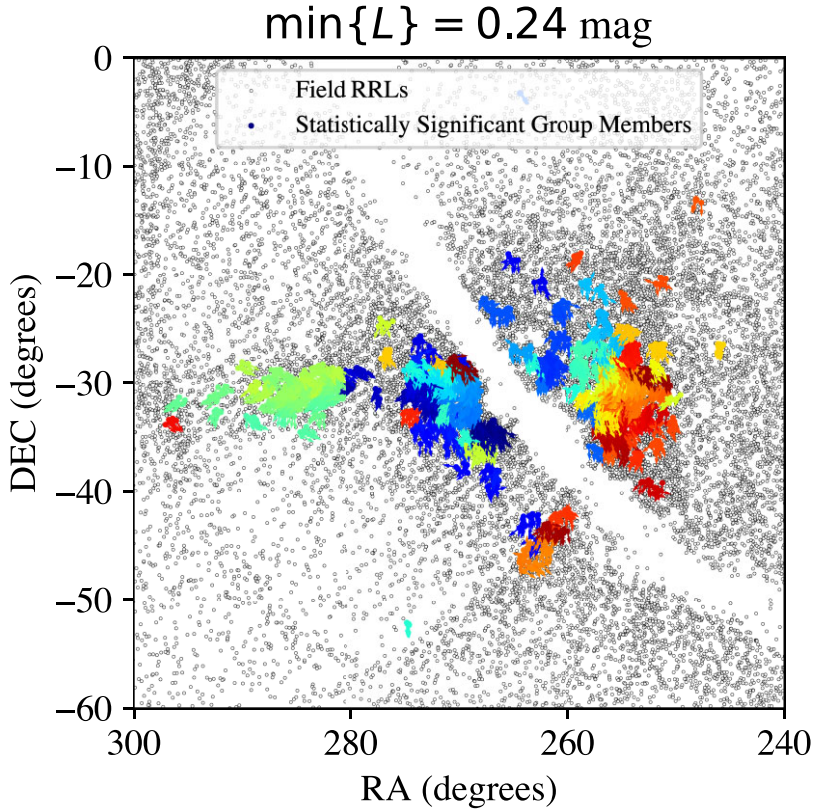


Figure 4. Hierarchical clustering at a scale ($\min\{L\} \simeq 0.24$ mag) that contains many RR Lyrae variables belonging to statistically significant groupings with correlated proper motions ($\tilde{X}_\theta \geq 0.98$, see equation 17). Variables colour-coded by their identified grouping are indicated by bigger markers to distinguish from field variables at this scale, and (normalized) the arrows indicate *Gaia* proper motion directions. The hierarchical clustering tree used to generate this figure contains all RR Lyrae variables in the displayed RA/DEC window, where subsequent figures use HEALPix subregions. The displayed region includes both the Galactic bulge (mostly blue and red groupings) and the Sagittarius Dwarf Spheroidal Galaxy (mostly green groupings), both of which contain many RR Lyrae variables and are, as a result, suitable for this illustration. Some of the groupings shown could be omitted from later figures, e.g. if their collective distance modulus spread indicates a separation distance in physical space of greater than $R_{\rm max} = 1$ kpc.

$$\frac{R}{[1\,\mathrm{kpc}]} = 10^{(\mu+\Delta\mu/2)/5-2} - 10^{(\mu-\Delta\mu/2)/5-2}, \tag{18}$$

$$\Delta\mu(\mu,\, R{=}R_{\rm max}) = 10\log_{10}\left(\frac{1}{2}\left[10^{2-\mu/5}+\sqrt{10^{2(2-\mu/5)}+4}\right]\right). \tag{19}$$

We collect all potentially relevant groupings, regardless of size, but then discard potentially spurious groupings whose radius in distance modulus units indicates a physical size greater than $R = R_{\rm max}$.

Once the RR Lyrae streams candidates were properly identified, we applied an ellipsoid fitting algorithm (Bazhin 2019) such that the principal axis vectors $\{\mathbf{e}_k\}$ of the best-fitting ellipsoid are provided. Given our interest in GCs and stellar streams, we applied the following scheme to the set of principal axis lengths $\{e_k\}$: if the eccentricity $e \equiv \sqrt{1 - \min(\{e_k\})/\max(\{e_k\})}$ was $\leq$0.2, the grouping was classified as a GC-like grouping (see Harris & Racine 1979 and Staneva, Spassova & Golev 1996 for justification), and if $e \geq 0.7$ it was identified as a stellar stream-like grouping (Martin & Jin 2010). Groupings with eccentricities in the intermediate range could in principle be admitted, but classified with the designation 'other'; in all cases, the grouping radius is defined as $R_{\rm grouping} \equiv \max\{e_k\}$.

Most of the identified groupings are classified as stellar streams within this paradigm. If there are outliers (which are permissible but unlikely with average linkage) or the number of grouped RR Lyraes is small, this eccentricity calculation would be susceptible to unacceptably high variations if new RR Lyraes were included. In such cases, this eccentricity computation should be thought of as a first-order estimate. It is beyond the scope of this study, but in instances where the internal structure of a grouping is important, ellipsoid fitting is ill-advised. A minimum spanning tree (Kruskal 1956), for example, could be used to further analyse the integrity of the grouping instead.

2.4 Finding Palomar 5 with *Gaia* RR Lyrae variables

A collection of RR Lyraes from the *Gaia* DR2 and Pan-STARRS1 (Sesar et al. 2017) catalogues were recently used to analyse the kinematic properties of the Palomar 5 stellar stream (Price-Whelan et al. 2019), a system whose properties make it suitable for analyses of the Galactic potential (e.g. Bovy et al. 2016; Starkman, Bovy & Webb 2020). The publicly available data set included the 3300 RR Lyrae variables found in the appropriate section of the sky plane, but in our pre-processing we only retained the 2348 variables of type RRab. A set of stellar stream model parameters were inferred with Markov chain Monte Carlo posterior sampling, and each of the catalogue variables were given a posterior membership probability.

Fig. 5 shows the one grouping identified in the refined *Gaia* catalogue via hierarchical clustering, and it is consistent with the Palomar 5 GC. This grouping contained 15 RRab variables, eight of which were within the GC's reported Jacobi radius. We used the

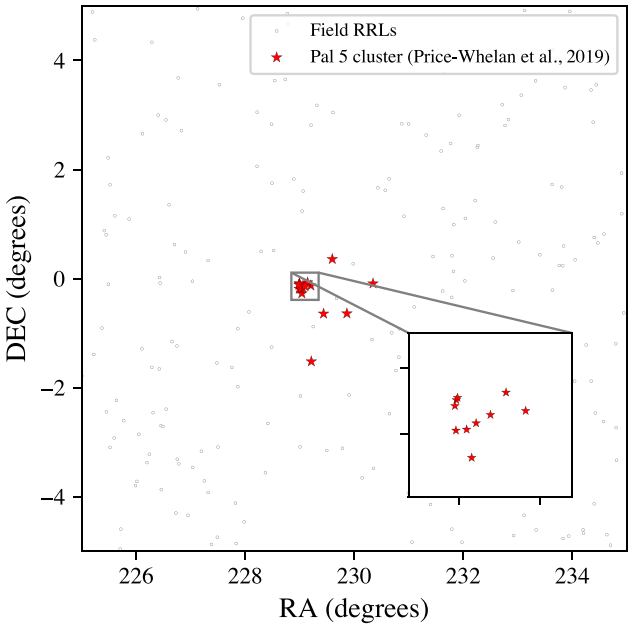


Figure 5. An RR Lyrae grouping identified by the clustering algorithm described in Section 2.3 that is consistent with the Palomar 5 GC. In comparing with a map of nearby RR Lyrae variables colour-coded by membership probability (Price-Whelan et al. 2019, fig. 3), we see that the inset RR Lyrae variables are within Pal 5's Jacobi radius (see Appendix A for a definition).

reported physical distances (in kiloparsecs) to infer each variable's distance modulus via equation (3). While the grouping has an eccentricity $e_{\text{grouping}} \simeq 0.9$, which qualifies as a stellar stream in our paradigm, this is attributable to the external variables included in the grouping and the general nature of average linkage. The grouping has a median distance modulus of $\mu_{\text{grouping}} \simeq 16.6$ mag, which is roughly consistent with Palomar 5. 13 of the 15 variables identified by our algorithm were awarded an MCMC membership probability of $p \gtrsim 0.5$; this, along with the absence of false positives, provides a reliable cross-check of our cluster identification routine against a known substructure.

3 RESULTS

In order to find archaeologically relevant substructures from the *Gaia* data via hierarchical clustering, we must make an informed decision of how to allocate the variables appropriately [e.g. consider employing random subsampling (Khoperskov et al. 2020) or breaking up the catalogue into smaller trees]. If there is a substructure of $N_\star$ RR Lyrae variables found in a catalogue of size $N_{\text{catalogue}}$ (assuming the described grouping cultivation routine is guaranteed to find it), then the probability that it is found within a randomly selected subset of size N_{subset} is

$$P(N_\star) \simeq \frac{N_{\text{subset}}!}{N_{\text{catalogue}}!} \frac{(N_{\text{catalogue}} - N_\star)!}{(N_{\text{subset}} - N_\star)!}. \tag{20}$$

The detection probability $P(N_\star)$ decreases dramatically as N_{subset} decreases, which is a quantitative motivation for choosing as large a catalogue of RR Lyrae variables as possible. However, the tree cutting process is not guaranteed to be well suited for an all-sky plane catalogue; indeed, the results shown in Figs 4 and 5 were determined using hierarchical clustering trees comprised of RR Lyrae variables from a comparatively small patch of the sky plane.

These conflicting attributes of hierarchical clustering trees motivate the following choice: we created 48 hierarchical clustering trees covering distinct HEALPix[3] subregions of equal sizes ($A_{\text{subregion}} \simeq 859$ deg^2) on the sky plane, following the RING ordering convention. The union of each tree's subpopulations is the entire catalogue of 91 234 variables shown in Fig. 2. We iterate through each of the 48 trees, collecting substructures with the desired proper motion and physical size attributes into a data base.

The resulting data base contained 32 candidate substructures, broken up into Tables 1 and B1, with a preference towards deeper analyses of the first group of substructures (shown in Fig. 6). The separation of substructure candidates took the following attributes into consideration: number of RR Lyrae variables, physical size, eccentricity, and proximity to the nearest GC. If a substructure candidate had a larger physical size, it likely means that the constituent RR Lyrae variables are less likely to be associated with any parent GC system. Substructures with IDs 2–4, for example, are RR Lyrae variable sets with, presumably, large intersections. Our stated preference for analysing the substructures in Table 1 is, admittedly, a subjective one.

Upon inspection, there are three substructure classes in Table 1: substructures within a kiloparsec of a Milky Way GC (IDs 3, 13, 15, 17, 22), substructures whose connections to any particular GC warrants further investigation (0, 9, 15, 16, 19, 24, 25, 27, 29, 30), and substructures that may be independent of the Milky Way GC population (28, 31). GCs with a first class substructure less than 1 kiloparsec away are highlighted in Table 2. Substructures 3 and 17 are likely related to the Fimbulthul (Ibata et al. 2019a) and Fjörm (Ibata, Malhan & Martin 2019b) streams, while substructures 13, 15, and 22 are in the vicinity of known GC tidal tails (e.g. Piatti & Carballo-Bello 2020; Ibata et al. 2021).

The second class of substructures have less straightforward connections to neighbouring GCs, and this is, in part, due to increasingly unforgiving distance modulus spreads. Substructures 0, 9, 15, 16, 25, and 27 are likely part of their listed nearest GC systems, but the more distant substructure members causing high substructure eccentricities (or distance modulus uncertainties) likely caused an identified substructure centre separation $d_{\text{substruct}\to\text{GC}} \gtrsim \Delta\mu(\tilde{\mu}, R = 1\,\text{kpc})$.

Substructures 19 and 24 ($\tilde{\mu}_{19} = 15.13$, $\tilde{\mu}_{24} = 15.62$) are closest to NGC 6266 ($\mu \simeq 14.2$) and NGC 6316 ($\mu \simeq 15.09$), respectively, despite appearing to be virtually identical in Fig. 6. The high number density of RR Lyrae variables in this region, as well as the number of nearby Milky Way GCs, makes the attribution of these substructures to any particular globular custer of stellar stream beyond the scope of this study. Substructures 29/30 are nearly redundant and consistent with the IC 4499 GC, a GC with a robust RR Lyrae population of type ab (Ferraro et al. 1995) and large tidal radius (Walker et al. 2011). While the sky plane configuration of substructures 29/30 are encouraging for making connections to IC 4499, their distance moduli ($\tilde{\mu}_{29} = 17.03$, $\tilde{\mu}_{30} = 17.14$) are larger than available estimates for the distance to IC 4499 (e.g. $\mu \simeq 16.5$, Storm 2004).

Substructure 28 ($\tilde{\mu} = 16.58$) is closest to IC 1257 ($\mu \simeq 17.0$, Harris et al. 1997), a small GC in the mid-halo region of the Milky Way. The presence of the NGC 6402 and NGC 6366 GCs, whose distance moduli are $\mu \simeq 15$, introduces confusion. This substructure's association with any of these three GCs is dubious, as the distances are inconsistent by several kiloparsecs. Additionally, none of these

[3] https://healpix.jpl.nasa.gov/

Table 1. Bulk attributes (e.g. median declination) of RR Lyrae substructures identified using the methods delineated in Section 2 and displayed in Fig. 6, sorted by distance modulus. The reported radius is converted into physical units using equation (18). The eccentricities of the identified substructures suggest pronounced elongation consistent with stellar streams. Substructures whose distance modulus separation from the nearest Milky Way GC (Harris 2010) is less than $\Delta\mu(\tilde{\mu}, R = R_{\max})$ (see equation 19) are more likely to be associated with said GCs, while the others warrant further investigation.

ID	$\tilde{\mu}$ [mag]	$\tilde{\alpha}$ [deg]	$\tilde{\delta}$ [deg]	R [pc]	$N_\star$	Eccentricity	Nearest GC	$d_{\text{substruct}\to\text{GC}}$ [mag]	$\Delta\mu(\tilde{\mu}, R = 1\,\text{kpc})$ [mag]
0	12.64	245.9	−26.5	225.11	29	0.996877	NGC 6121	0.93	0.64
3	13.81	201.6	−47.4	749.46	16	0.983881	NGC 5139	0.23	0.38
9	14.25	154.4	−46.4	401.87	22	0.998333	NGC 3201	0.80	0.31
13	14.39	229.6	2.1	818.17	27	0.998918	NGC 5904	0.02	0.29
15	14.58	263.0	−67.1	922.35	22	0.915398	NGC 6362	0.18	0.26
16	14.78	248.1	−13.1	160.32	20	0.967803	NGC 6171	0.75	0.24
17	15.00	189.9	−26.7	484.83	18	0.968888	NGC 4590	0.06	0.22
19	15.13	255.3	−30.1	274.61	26	0.942493	NGC 6266	0.97	0.20
22	15.39	78.5	−40.1	228.64	15	0.997036	NGC 1851	0.03	0.18
24	15.62	255.3	−30.1	454.92	20	0.997226	NGC 6316	1.17	0.16
25	16.30	313.4	−12.5	526.09	26	0.995423	NGC 6981	0.15	0.12
27	16.32	308.6	7.4	382.78	26	0.981308	NGC 6934	0.35	0.12
28	16.58	264.4	−3.2	221.74	15	0.981297	IC 1257	1.43	0.10
29	17.03	225.1	−82.2	737.69	21	0.998445	IC 4499	0.66	0.09
30	17.14	225.1	−82.2	955.57	21	0.987165	IC 4499	0.77	0.08
31	18.71	96.8	−70.1	741.74	15	0.937961	E 3	5.73	0.04

GCs have known tidal tails (Piatti & Carballo-Bello 2020; Ibata et al. 2021). Therefore, this substructure is an appealing candidate for future studies.

The substructure with the largest distance modulus, substructure 31 ($\tilde{\mu}_{31} = 18.71$), is only separated from the LMC by a few degrees on the sky plane and not likely to be associated with the E 3 GC (as suggested in Table 2). With a physical distance of $d_{\odot\to\text{substruct}} \simeq 55.2$ kpc, substructure 31 is very likely a member of the Magellanic system. There is evidence that the Magellanic system was previously a triplet of dwarf galaxies before the LMC accreted one of its siblings (e.g. Armstrong & Bekki 2018; Mucciarelli et al. 2021), and two kinematically distinct GC populations (Piatti, Alfaro & Cantat-Gaudin 2019b) suggest that the Magellanic system is a good place to search for artefacts of the hierarchical galaxy formation process. A runaway star cluster east of the LMC was recently discovered in a series of recent deep imaging surveys (Piatti, Salinas & Grebel 2019a; Piatti 2021) near substructure 31; the reported age of the star cluster ($0.89^{+0.11}_{-0.10}$ Gyr), however, is inconsistent with a substructure of Population II stars. An analysis of tile 5-9 in the VMC survey's RR Lyrae variable catalogue (Cusano et al. 2021) is likely needed to confirm the existence of this substructure and its properties.

4 DISCUSSION

14 RR Lyrae variable streams have been identified from the Catalina survey at $>3.5\sigma$ confidence (Mateu, Read & Kawata 2018), several of which are located near Milky Way GCs. We do not claim that the groupings shown in Fig. 6 rise to this level of precision, as there is evidence that spectroscopy is required to confirm Galactic stellar stream candidates even in the *Gaia* era (Jean-Baptiste et al. 2017). Instead, we will proceed with a discussion of the identified substructures, and their local environments, knowing that further investigation is needed. Follow-up studies would not only need to confirm each candidate grouping's legitimacy (likely through chemical abundance matching and an exploration of the entire phase space), but also to determine the entire scope of the candidate substructure beyond its population of RR Lyrae variables.

We determined each candidate grouping's nearest GC neighbour in the (x, y, z) space defined in Section 2.3; the results are shown in Table 2. Eight of the 15 GCs listed in this table were accreted during merger events (Kruijssen et al. 2020) and seven have been noted in the literature as having extratidal features in their outermost regions (Piatti & Carballo-Bello 2020, Ibata et al. 2021, and references therein). Additionally, *Gaia* EDR3 RR Lyrae variables have escaped from two of the 15 GCs (NGC 5904 and NGC 1851) according to a recent study of proper motions and colour–magnitude diagrams (Abbas, Grebel & Simunovic 2021). If there are new connections to be made between Milky Way GCs and the Galaxy's RR Lyrae population via substructure identification, we suggest focusing on the following substructures where there is no previously known tidal tail despite being a potential progenitor traced back to a Galactic merger event: 0, 25, 29, 30. The `Via Machinae` algorithm (Shih et al. 2022), for example, is an unsupervised machine learning routine designed to detect stellar streams, and would likely be a suitable method for carrying out such a follow-up study.

The nuclear star clusters that remain from four of these merger events were recently identified using chemo-kinematic information (Pfeffer et al. 2021), and our routine identified RR Lyrae substructures near two of them: NGC 5139 (ω Centauri), reportedly the remnant of the *Gaia*-Enceladus merger, and NGC 6934, the reported remnant of the Helmi streams merger (although NGC 6934's status as the merger remnant is less certain). Categorizing accreted GCs by their parent merger events, and even compartmentalizing the merger events themselves, is an ongoing area of research. A recent study of 23 known Galactic stellar streams, for example, demonstrated that clustering in orbital phase space is a fruitful method of identifying their sources (Bonaca et al. 2021). Many of these streams are extended for $\gtrsim 100^\circ$ on the sky; our routine's inability to identify structures of this kind further motivates our focus on RR Lyrae substructures that may be components of stream cores in the neighbourhood of Milky Way GCs.

One of the benefits of employing hierarchical clustering is that we can select a context-dependent distance metric between objects. While it is possible that the intersection of the *Gaia* DR2 RR Lyrae variable catalogue and the APOGEE-2 survey's collection

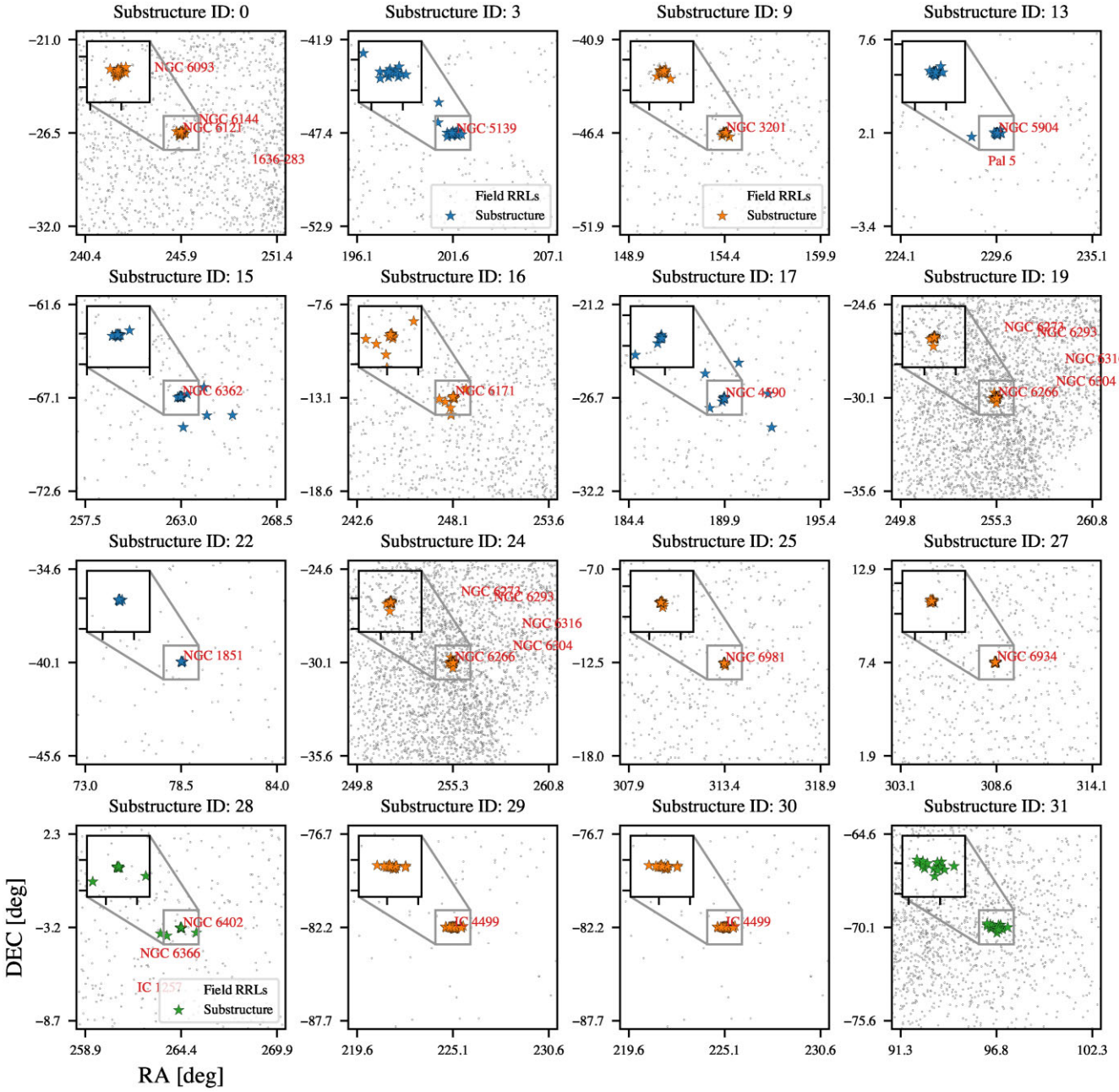


Figure 6. Substructures listed in Table 1, colour-coded by substructure class as defined in Section 3. Field RR Lyrae variables from the cultivated catalogue and local GCs on the sky plane are included. A $2^\circ \times 2^\circ$ inset in each panel provides a zoomed-in view of each of the substructure cores, most of which are consistent with Milky Way GCs.

of absorption spectra[4] is prohibitively small, a more sophisticated metric that incorporates stellar motions and chemical abundances is worth considering in future work. A potential improvement to this methodology would include a Milky Way phase space and metallicity vector $\boldsymbol{\chi} = (\mathbf{x}, \mathbf{v}, [\mathrm{Fe/H}])$, as well as a metric D_{modified} in a chemo-kinematic space dependent on the Galactic integrals of motion E_{tot} and $\mathbf{L}$:

$$\begin{aligned} D_{\mathrm{modified}}(\boldsymbol{\chi}, \boldsymbol{\chi}')^2 \equiv\ & \beta_0 (E_{\mathrm{tot}} - E'_{\mathrm{tot}})^2 \\ & + \beta_1 (L_z - L'_z)^2 \\ & + \beta_2 (L_\perp - L'_\perp)^2 \\ & + \beta_3 ([\mathrm{Fe/H}] - [\mathrm{Fe/H}]')^2, \end{aligned} \tag{21}$$

where the total energy is $E_{\mathrm{tot}} = (\mathbf{v} \cdot \mathbf{v})/2 + \Phi(\mathbf{x})$, the angular momentum is $\mathbf{L} = \mathbf{x} \times \mathbf{v}$, and $\{\beta_i\}$ are scaling factors that ensure each term has the appropriate units and weighted importance. Further investigation is needed to determine if this is a suitable metric or

[4]From the Sloan Digital Sky Survey website (https://www.sdss.org/surveys/apogee-2/): 'The second generation of the Apache Point Observatory Galaxy Evolution Experiment (APOGEE-2) observes the 'archaeological' record embedded in hundreds of thousands of stars to explore the assembly history and evolution of the Milky Way Galaxy.' This data set contains information derived from spectroscopic measurements in the near-infrared.

Table 2. Milky Way GCs (with reported distance moduli and Milky Way coordinates, Harris 2010) near an identified RR Lyrae substructure presented in Section 3. Each GC's potential association with a progenitor merger event (Kruijssen et al. 2020), known extratidal features (Piatti & Carballo-Bello 2020), and well-studied Galactic stellar streams (Bonaca et al. 2021), if applicable, are provided. [G1 denotes a symmetric tidal tail, G2 a feature outside of the Jacobi radius that is not necessarily a tidal tail, G3 no signature of extended structure, and SF a long tidal tail discovered using the `STREAMFINDER` algorithm on *Gaia* DR2/EDR3 data (Ibata et al. 2021)]. This classification scheme is only applied to previously known substructures and is independent of the results presented in this paper.) Highlighted GC IDs indicate an instance where an RR Lyrae substructure was identified at a physical distance of less than 1 kpc.

GC ID	Substructure ID	μ_{GC} [mag]	α [deg]	δ [deg]	Potential progenitor	Tidal tail	Stream name
NGC 6121	0	11.71	245.90	−25.47	Kraken		
NGC 3201	9	13.45	154.40	−45.59	Sequoia/*Gaia*-Enceladus	G2/SF	
NGC 5139	3	13.58	201.70	−46.52	*Gaia*-Enceladus/Sequoia	G1/SF	Fimbulthul
NGC 6171	16	14.03	248.13	−12.95			
NGC 6266	19	14.16	255.30	−29.89		G2	
NGC 5904	13	14.38	229.64	2.08	Helmi streams/*Gaia*-Enceladus	G1/SF	
NGC 6362	15	14.40	262.98	−66.95		G2	
E 3	31	14.54	140.24	−76.72			
NGC 4590	17	15.06	189.87	−25.26	Helmi streams	G1/SF	Fjörm
NGC 6316	24	15.09	259.16	−28.14			
NGC 1851	22	15.41	78.53	−39.95	*Gaia*-Enceladus	G1/SF	
NGC 6934	27	15.97	308.55	7.40			
NGC 6981	25	16.15	313.37	−11.46	Helmi streams		
IC 4499	29, 30	16.37	225.08	−81.79	Sequoia		
IC 1257	28	16.99	261.79	−6.91			

if different vector components should be introduced. Additionally, it is entirely possible that a large RR Lyrae data set suitable for this modified metric on a Galactic scale is not yet available. In the future, infrared data collected by the Nancy Grace Roman Space Telescope could be ideal for this type of analysis. The period–luminosity relation is, strictly speaking, only valid in the infrared (see Section 1), and this relation is what makes RR Lyrae variable stars such reliable standard candles. Roman Telescope IR photometry and spectroscopic surveys could in principle be combined for more accurate RR Lyrae distance moduli (equation 2). Additionally, this would provide enough chemo-kinematic information such that the modified metric (equation 21) might be tested and applied. The existence of a Galactic archaeology consortium and data analysis pipeline, as presented in Ness et al. (2019), would certainly help make a study of this kind possible.

An agglomerative, hierarchical clustering algorithm utilizing average linkage has time complexity $O(n^2 \log_2 n)$ (Manning, Raghavan & Schütze 2008), so applying particle-based clustering of this kind is probably ill-advised for samples bigger than this one. Density-based hierarchical clustering, with time complexity $O(n[\log_2 n]^3)$, has been successfully employed in astronomical contexts (Sharma & Johnston 2009; Elahi 2013; Sanderson, Helmi & Hogg 2015; Sanderson et al. 2017). Future work could be devoted to applying such an algorithm to a larger catalogue such that the relevant instrument's selection function might be mitigated via random sampling.

Clustering algorithms are some of the most popular unsupervised learning processes; given the nature of astronomical data sets (namely, RR Lyrae variables do not have name tags declaring to which GC or dwarf galaxy progenitor it belongs) it is natural to employ such processes. However, it may be possible to identify structures of interest using supervised learning once a sufficient number of training examples become available via observations or simulations. For example, a random forest classifier (Breiman 2001) could be employed, where labelled inputs (i.e. list of RR Lyraes known to belong to a particular GC) pass through a forest of decision trees whose parameters are randomly chosen from the input's data. The advantage of a random forest in this context is the incorporation of quantities not captured by an image, like metallicities or phase space coordinates. In the event such a random forest regimen is insufficiently accurate, implementing gradient boosting via XGBoost (Chen & Guestrin 2016) would be a possible improvement.

5 CONCLUSION

We have presented an analysis of the *Gaia* mission's RR Lyrae variable catalogue; the basis for this study was the repeated application of an agglomerative, hierarchical clustering algorithm to subsets of the variable star catalogue. The uncertainty in computed distance as a function of measurement errors in absolute/apparent magnitude and associated extinction is considered, and we determined that the largest driver of distance modulus uncertainties results from large values of δm_V. Once each RR Lyrae variable's 3D spherical coordinates were compiled (with the distance modulus used as a proxy for physical distance), we computed the condensed distance matrices for 48 catalogue subsets partitioned by HEALPix subregion. These matrices were used to create hierarchical clustering trees (with average linkage) containing archaeologically relevant substructures of RR Lyrae variables. The hierarchical clustering trees contain many nearly redundant groupings at neighbouring clustering scales, so we proceeded in analysing select scales with a local maximum of statistically significant groupings. The potentially interesting groupings were then compiled and analysed.

Our results suggest that hierarchical clustering trees that use average linkage contain primarily stream extensions around GCs in a variety of Galactic environments. Substructures 0, 25, 29, and 30 are good candidates for making new connections to GCs important to the formation of the Milky Way; substructure 28 lies along the sightline near several known GCs, but has a distance modulus that suggests it is independent of these systems; substructure 31 is a possibly previously unknown satellite of the LMC. Follow-up studies would benefit from an exploration of the entire phase space (e.g. Gaia EDR3 and DR3, Gaia Collaboration 2021), as well as a sufficiently comprehensive mapping of the Milky Way that fills in any coverage gaps (as the Rubin Observatory is expected to provide in the coming years, Najita et al. 2016). Applying hierarchical clustering to three-dimensional data and validating with proper motions, however, is indeed effective in identifying groupings of RR Lyrae variable stars,

one of the undoubtedly reliable tracers of Galactic structure available in the cosmos.

Software: NUMPY (Oliphant 2007), SCIPY (Virtanen et al. 2020), MATPLOTLIB (Hunter 2007), PANDAS (McKinney 2010), ASTROPY (Price-Whelan et al. 2018).

ACKNOWLEDGEMENTS

This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

The authors wish to thank Anthony Brown, Marta Reina-Campos, and the anonymous referee for insightful comments that helped improve the manuscript.

BTC, KDS, and RM were supported by the Summer Research Program at MIT Lincoln Laboratory. BTC would like to thank the MIT Lincoln Laboratory staff, whose support during the internship made this work possible.

Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

DATA AVAILABILITY STATEMENT

The data underlying this article are publicly available; they can be found using the citations and footnotes provided throughout the article. The article's online supplementary material includes a collection of this public data in a form that is well suited for this study.

REFERENCES

Abbas M., Grebel E. K., Simunovic M., 2021, ApJ, 915, 49
Akaike H., 1998, Information Theory and an Extension of the Maximum Likelihood Principle. Springer, New York, NY, p. 199
Armstrong B., Bekki K., 2018, MNRAS, 480, L141
Baumgardt H., Hilker M., 2018, MNRAS, 478, 1520
Bazhin A., 2019, ellipsoid fit python. https://github.com/aleksandrbazhin/ellipsoid_fit_python
Belokurov V., Deason A. J., Erkal D., Koposov S. E., Carballo-Bello J. A., Smith M. C., Jethwa P., Navarrete C., 2019, MNRAS, 488, L47
Binney J., Tremaine S., 2008, Galactic Dynamics, 2nd edn. Princeton University Press, Princeton, NJ
Bonaca A. et al., 2021, ApJ, 909, L26
Bovy J., 2015, ApJS, 216, 29
Bovy J., Bahmanyar A., Fritz T. K., Kallivayalil N., 2016, ApJ, 833, 31
Breiman L., 2001, Mach. Learn., 45, 5
Bullock J. S., Johnston K. V., 2005, ApJ, 635, 931
Catelan M., Pritzl B. J., Smith H. A., 2004, ApJS, 154, 633
Chaboyer B., 1999, Globular Cluster Distance Determinations. Springer, Netherlands, Dordrecht, p. 111
Chen T., Guestrin C., 2016, preprint(arXiv:1603.02754)
Clauset A., Moore C., Newman M. E. J., 2008, Nature, 453, 98
Clement C. M. et al., 2001, AJ, 122, 2587
Clementini G. et al., 2019, A&A, 622, A60
Cusano F. et al., 2021, MNRAS, 504, 1
De Silva G. M. et al., 2015, MNRAS, 449, 2604
Duffau S., Zinn R., Vivas A. K., Carraro G., Méndez R. A., Winnick R., Gallart C., 2006, ApJ, 636, L97
Elahi P. J., 2013, Astrophysics Source Code Library, record ascl:1306.009
Everitt B., Landau S., Leese M., Stahl D., 2011, Cluster Analysis, 5th edn. Wiley, Hoboken, NJ
Ferraro I., Ferraro F. R., Pecci F. F., Corsi C. E., Buonanno R., 1995, MNRAS, 275, 1057
Gaia Collaboration, 2021, A&A, 649, A1
Gordon A. D., 1987, J. R. Stat. Soc. A, 150, 119
Gower J. C., Ross G. J. S., 1969, Appl. Stat., 18, 54
Green G. M., 2018, J. Open Source Softw., 3, 695
Green G. M., Schlafly E., Zucker C., Speagle J. S., Finkbeiner D., 2019, ApJ, 887, 93
Harris W. E., 2010, preprint (arXiv:1012.3224)
Harris W. E., Racine R., 1979, ARA&A, 17, 241
Harris W. E., Phelps R. L., Madore B. F., Pevunova O., Skiff, Brian A. Crute C., Wilson B., Archinal B. A., 1997, AJ, 113, 688
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ibata R. A., Bellazzini M., Malhan K., Martin N., Bianchini P., 2019a, Nat. Astron., 3, 667
Ibata R. A., Malhan K., Martin N. F., 2019b, ApJ, 872, 152
Ibata R. et al., 2021, ApJ, 914, 123
Iorio G., Belokurov V., 2019, MNRAS, 482, 3868
Ivezić Ž., Connelly A. J., Vand erPlas J. T., Gray A., 2014, Statistics, Data Mining, and Machine Learning in Astronomy. Princeton University Press, Princeton, NJ
Jänes J., Pelupessy I., Portegies Zwart S., 2014, A&A, 570, A20
Jean-Baptiste I., Di Matteo P., Haywood M., Gómez A., Montuori M., Combes F., Semelin B., 2017, A&A, 604, A106
Johnston K. V., Bullock J. S., Sharma S., Font A., Robertson B. E., Leitner S. N., 2008, ApJ, 689, 936
Khoperskov S., Gerhard O., Di Matteo P., Haywood M., Katz D., Khrapov S., Khoperskov A., Arnaboldi M., 2020, A&A, 634, L8
Kissler-Patig M., 1999, in Carral P., Cepa J., eds, ASP Conf. Ser. Vol. 163, Star Formation in Early Type Galaxies. Astron. Soc. Pac., San Francisco, p. 184
Kolenberg K., Fossati L., Shulyak D., Pikall H., Barnes T. G., Kochukhov O., Tsymbal V., 2010, A&A, 519, A64
Kruijssen J. M. D. et al., 2020, MNRAS, 498, 2472
Kruskal J. B., 1956, Proc. Am. Math. Soc., 7, 48
Langfelder P., Zhang B., Horvath S., 2007, Bioinformatics, 24, 719
Lindegren L. et al., 2021, A&A, 649, A2
Maciejewski M., Colombi S., Springel V., Alard C., Bouchet F. R., 2009, MNRAS, 396, 1329
Maeder A., 2009, Physics, Formation and Evolution of Rotating Stars. Springer Nature, Berlin, Germany
Malhan K., Ibata R. A., Carlberg R. G., Bellazzini M., Famaey B., Martin N. F., 2019, ApJ, 886, L7
Manning C. D., Raghavan P., Schütze H., 2008, Introduction to Information Retrieval. Cambridge Univ. Press, New York, NY, USA
Martin N. F., Jin S., 2010, ApJ, 721, 1333
Mateu C., Read J. I., Kawata D., 2018, MNRAS, 474, 4112
McKinney W., 2010, in van der Walt S., Millman J., eds, Proceedings of the 9th Python in Science Conference. p. 51
Mucciarelli A., Massari D., Minelli A., Romano D., Bellazzini M., Ferraro F. R., Matteucci F., Origlia L., 2021, Nat. Astron., 5, 1247
Najita J. et al., 2016, preprint (arXiv:1610.01661)
Ness M. et al., 2019, preprint (arXiv:1907.05422)
Oliphant T. E., 2007, Comput. Sci. Eng., 9, 10
Pearson S., Price-Whelan A. M., Johnston K. V., 2017, Nat. Astron., 1, 633
Peebles P. J. E., 1974, A&A, 32, 197
Pelupessy F. I., Jänes J., Portegies Zwart S., 2012, New Astron., 17, 711
Pelupessy F. I., van Elteren A., de Vries N., McMillan S. L. W., Drost N., Portegies Zwart S. F., 2013, A&A, 557, A84
Pfeffer J., Lardo C., Bastian N., Saracino S., Kamann S., 2021, MNRAS, 500, 2514
Piatti A. E., 2021, A&A, 647, A47
Piatti A. E., Carballo-Bello J. A., 2020, A&A, 637, L2

Piatti A. E., Salinas R., Grebel E. K., 2019a, MNRAS, 482, 980
Piatti A. E., Alfaro E. J., Cantat-Gaudin T., 2019b, MNRAS, 484, L19
Portegies Zwart S. F., 2009, ApJ, 696, L13
Portegies Zwart S., McMillan S., 2018, Astrophysical Recipes; The art of AMUSE. IOP Publishing, Bristol, UK
Portegies Zwart S. et al., 2009, New Astron., 14, 369
Portegies Zwart S., McMillan S. L. W., van Elteren E., Pelupessy I., de Vries N., 2013, Comput. Phys. Commun., 184, 456
Press W. H., Schechter P., 1974, ApJ, 187, 425
Preston G. W., 1964, ARA&A, 2, 23
Price-Whelan A. M. et al., 2018, AJ, 156, 123
Price-Whelan A. M., Mateu C., Iorio G., Pearson S., Bonaca A., Belokurov V., 2019, AJ, 158, 223
Sandage A., Cacciari C., 1990, ApJ, 350, 645
Sanderson R. E., Helmi A., Hogg D. W., 2015, ApJ, 801, 98
Sanderson R. E., Secunda A., Johnston K. V., Bochanski J. J., 2017, MNRAS, 470, 5014
Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103
Searle L., Zinn R., 1978, ApJ, 225, 357
Sesar B. et al., 2013, AJ, 146, 21
Sesar B., Hernitschek N., Dierickx M. I. P., Fardal M. A., Rix H.-W., 2017, ApJ, 844, L4
Sharma S., Johnston K. V., 2009, ApJ, 703, 1061
Shih D., Buckley M. R., Necib L., Tamanas J., 2022, MNRAS, 509, 5992
Smith H. A., 1995, Cambridge Astrophysics Series Vol. 27. Cambridge University Press, Cambridge, UK
Staneva A., Spassova N., Golev V., 1996, A&AS, 116, 447
Starkman N., Bovy J., Webb J. J., 2020, MNRAS, 493, 4978
Storm J., 2004, A&A, 415, 987
Vasiliev E., 2019, MNRAS, 484, 2832
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Vivas A. K., Zinn R., Farmer J., Duffau S., Ping Y., 2016, ApJ, 831, 165
Walker A. R. et al., 2011, MNRAS, 415, 643
Watkins L. L. et al., 2009, MNRAS, 398, 1757
Zinn R., 1985, ApJ, 293, 424

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

RRL_database_from_gaia.csv

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: MOCK RR LYRAE STREAMS FROM AMUSE SIMULATION

In order to refine our RR Lyrae stream identification routine via an informed choice of $\tilde{X}_{\theta,\min}$, $R_{\max}$ (defined in Section 2.3), we must create a labelled data set reflective of realistic Galactic dynamics and stellar populations. We select GCs NGC 362, NGC 2419, NGC 5466, Pal 5, and Pal 12 for further analysis using Baumgardt & Hilker (2018) and Vasiliev (2019).[5] Each of these GCs have reported extratidal features and associations with massive merger events in the Milky Way's formation history. This set has a diverse sampling of masses and galactocentric distances as well.

GCs are notoriously difficult to model computationally, as the collisional nature of the internal dynamics necessitates using specialized gravity solvers. These *N*-body codes often carry time complexities of the order of $O(n^2)$; for $n \sim 10^6$, this can get prohibitively expensive. It is beyond the scope of this study to simulate the tidal stripping of GCs, so we propose the following simplified model of RR Lyrae stream creation. At each spatial location of the *i*th GC, we create a particle of mass $M_{\mathrm{GC},i}$ and calculate the relevant Jacobi radius (Binney & Tremaine 2008):

$$r_{J,i} = |\mathbf{r}_i| \left(\frac{M_{\mathrm{GC},i}}{3M_{\mathrm{MW}}(\mathbf{r}_i)} \right)^{1/3}, \tag{A1}$$

where $\mathbf{r}_i$ is the position vector of the *i*th GC at the present epoch and $M_{\mathrm{MW}}(\mathbf{r}_i)$ is the enclosed Galactic mass at this location. This is approximately the boundary at which the GC and Galactic gravitational fields have equal influence. We then initialize 40 particles with an appropriate mass for RR Lyrae variables, $M = 0.65\,M_\odot$ (Kolenberg et al. 2010), and give them randomly distributed circular orbits about the GC particle with a semimajor axis equal to the Jacobi radius at initialization time. If these RR Lyrae-like particles maintain orbits consistent with their natal GC, we should still be able to identify the remnant core using the max (N_{RRL}) value from Sesar et al. (2013). We initialize each GC in velocity space using representative Milky Way orbits from the `GALPY` package (Bovy 2015).

A natural runtime choice for this simulation is the GC crossing time, $t_{\mathrm{cross},i} \simeq |\mathbf{r}_i|/|\mathbf{v}_i|$. An order-of-magnitude estimate for the crossing time can be found using the enclosed Milky Way mass and assuming that the described model is virialized:[6]

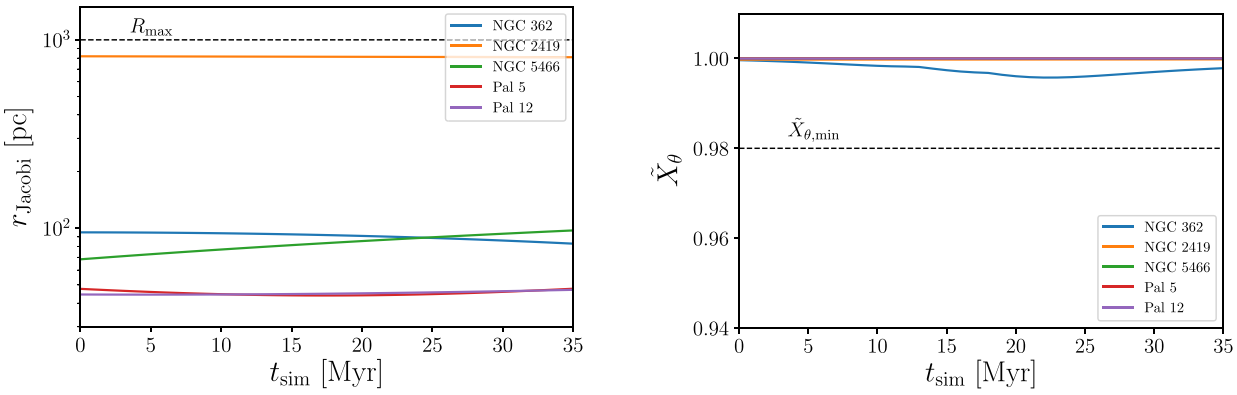
$$t_{\mathrm{cross},i} = \sqrt{\frac{|\mathbf{r}_i|^3}{GM_{\mathrm{MW}}(\mathbf{r}_i)}}. \tag{A2}$$

In order to ensure that the stream gravity solvers adequately conserve energy (i.e. a fractional error $\epsilon \lesssim 10^{-7}$ throughout the simulation), we use the time-step $\Delta t = 0.01$ Myr and a simulation time-scale $t_{\mathrm{end}} = 35$ Myr that is approximately the shortest crossing time (Pal 5). We simulate this system of massive particles with the aforementioned specifications using the AMUSE Python API (Portegies Zwart et al. 2009; Pelupessy et al. 2013; Portegies Zwart et al. 2013; Portegies Zwart & McMillan 2018); each stream is evolved forward in time using a symplectic (i.e. phase space conserving) integrator `Huayno` (Pelupessy, Jänes & Portegies Zwart 2012; Jänes, Pelupessy & Portegies Zwart 2014) and bridged with a Milky-Way-like (bar, disc, and bulge components) potential (Bovy 2015). Gravitational forces between streams are not considered, as they are negligible in comparison to background tidal forces.

Fig. A1 shows the values of $\tilde{X}_\theta$ and the Jacobi radius as a function of simulation time for each of the five streams, as well as our stream parameter choice: $(\tilde{X}_{\theta,\min}, R_{\max}) = (0.98, 1\ \mathrm{kpc})$. We approximated the proper motion vectors $(\mu_\alpha^*, \mu_\delta)$ using (v_y, v_z) from the simulation data, where *y* and *z* are part of the Galactic coordinate system defined in Section 2. The physical meaning of $\tilde{X}_{\theta,\min} = 0.98$ is that at least half of all stars in a retained grouping have proper motion vectors that deviate from the median grouping proper motion by 11.48°. This is a conservative threshold, as $\tilde{X}_\theta$ is very nearly equal to unity (i.e. proper motion deviations of the order of arcminutes rather than degrees) for all of the GC models during the simulation. Our choice of $R_{\max}$ ensures that two classes of substructure are retained: stellar streams extending an order-of-magnitude beyond the Jacobi radii of typical GCs, and atypical GCs like NGC 2419 that are more massive and further removed from the Galactic Center.

[5] https://github.com/GalacticDynamics-Oxford/GaiaTools/

[6] Some of the GCs we are analysing are extragalactic in nature, so this assumption should be used sparingly.



(a) The Jacobi radius (as defined in Equation (A1)) for each of the five GC models as a function of simulation time.

(b) The median alignments (as defined in Equation (16)) for each of the five GC models as a function of simulation time.

Figure A1. Motivation for the choice of ($\tilde{X}_{\theta,\rm min}$, $R_{\rm max}$) stream parameters in the substructure identification routine presented in Section 2.3. (a) The Jacobi radius (as defined in equation A1) for each of the five GC models as a function of simulation time. (b) The median alignments (as defined in equation 16) for each of the five GC models as a function of simulation time.

APPENDIX B: POTENTIALLY EXTRANEOUS SUBSTRUCTURE CANDIDATES

Substructures identified in the hierarchical clustering forest described in Sections 2 and 3 were partitioned into two groups, where the attributes of the second group are provided in Table B1. These substructures are related to those listed Table 1, but for the sake of brevity we focused our analysis on the ones listed there and include the remainder in this appendix.

An illustrative example of this organization is the retention of substructure 3 for Table 1, while listing substructures 2 and 4 are in Table B1. All three substructures have sky plane coordinates consistent with within 0.1° in both right ascension and declination with a similar number of RR Lyrae variables (15, 16, and 20, respectively). Each substructure is in the neighbourhood of the ω Centauri GC, a known remnant of a dwarf galaxy merger whose tidal stream, Fimbulthul, was recently discovered using *Gaia* DR2 data. The Fimbulthul stream is separated from its progenitor by nearly 20 deg on the sky plane and 1.5 kpc closer to Earth than ω Cen, thus making these candidate substructures unlikely members of the stream. Substructure 3 has the largest eccentricity, which means that if there is any connection to be made between ω Cen and the Fimbulthul stream via intermediate RR Lyrae variables (*N*-body simulations suggest a stream of stars connecting the two should be present, Ibata et al. 2019a), substructure 3 is the most promising candidate.

Table B1. Identified substructures where the constituent RR Lyrae variable sets are similar to the substructures listed in Table 1. The hierarchical clustering process, and the substructure identification routine described in Section 2, do not strictly prohibit nearly redundant groupings.

ID	$\tilde{\mu}$ [mag]	$\tilde{\alpha}$ [deg]	$\tilde{\delta}$ [deg]	R [pc]	$N_\star$	Eccentricity	Nearest GC	$d_{\rm substruct \to GC}$ [mag]	$\Delta\mu(\tilde{\mu}, R = 1\,{\rm kpc})$ [mag]
1	12.77	245.9	−26.5	138.48	19	0.994871	NGC 6121	1.06	0.60
2	13.81	201.6	−47.4	274.07	15	0.979803	NGC 5139	0.23	0.38
4	13.84	201.6	−47.4	264.15	20	0.887820	NGC 5139	0.26	0.37
5	13.99	154.4	−46.4	388.76	29	0.977655	NGC 3201	0.54	0.35
6	14.00	154.3	−46.4	239.44	23	0.985767	NGC 3201	0.56	0.34
7	14.01	154.4	−46.4	160.30	17	0.978757	NGC 3201	0.56	0.34
8	14.11	154.4	−46.4	126.51	19	0.946207	NGC 3201	0.66	0.33
10	14.38	229.6	2.1	779.15	26	0.998827	NGC 5904	0.01	0.29
11	14.38	229.6	2.1	779.15	26	0.998827	NGC 5904	0.01	0.29
12	14.39	229.6	2.1	276.70	23	0.986319	NGC 5904	0.02	0.29
14	14.58	263.0	−67.0	490.72	19	0.999308	NGC 6362	0.18	0.26
18	15.01	189.9	−26.7	240.83	16	0.992554	NGC 4590	0.06	0.22
20	15.13	255.3	−30.1	442.13	30	0.923307	NGC 6266	0.97	0.20
21	15.13	255.3	−30.1	258.84	27	0.948400	NGC 6266	0.97	0.20
23	15.44	255.3	−30.1	262.99	19	0.997485	NGC 6316	1.10	0.18
26	16.30	313.4	−12.5	278.98	32	0.932833	NGC 6981	0.15	0.12

This paper has been typeset from a TEX/LATEX file prepared by the author.